Title: In Situ Inference for Earth System Predictability

Author(s): Lawrence, Earl Christopher
Biswas, Ayan
Van Roekel, Luke
Urban, Nathan

Intended for: DOE White Paper

Issued: 2021-02-17

In Situ Inference for Earth System Predictability

Authors

Earl Lawrence (LANL), Ayan Biswas (LANL), Luke Van Roekel (LANL), Nathan Urban (Brookhaven NL)

Focal Area

Focal Area 3: Insight gleaned from complex simulated data using AI, big data analytics, and other advanced methods, including explainable AI and physics- or knowledge-guided AI.

Science Challenge

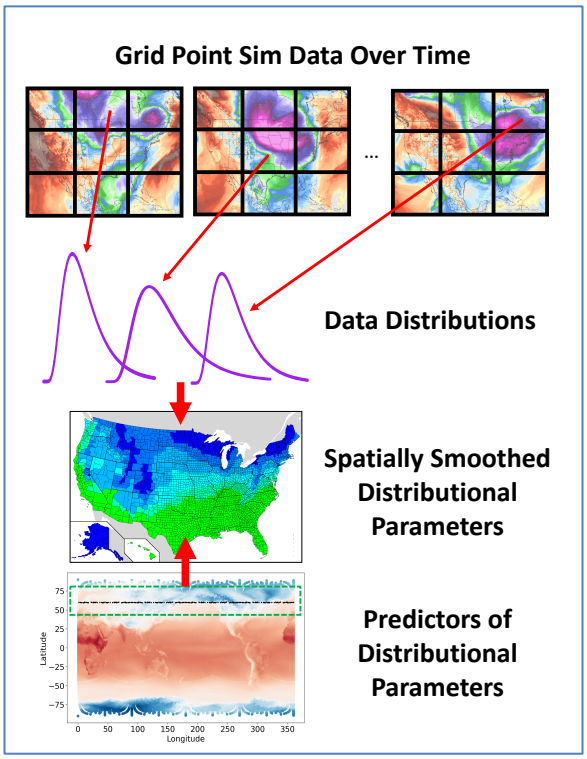

Fig 1: A Bayesian hierarchical spatial regression model. Precipitation at each simulation grid cell is modeled with an extreme value distribution who parameters are spatially smoothed functionally dependent on variables at other spatiotemporal locations (e.g. temp and wind speed at high latitudes).

An understanding of future evolution in precipitation extremes is critical to numerous DOE mission questions. Extreme events are by nature short time-scale events that are difficult to diagnose in available model data. Accurate modeling of extreme events necessarily requires high spatial resolution at the storm scale locally. However, the environment in which storms grow is dependent on global, remote, processes. These complex spatiotemporal relationships are impossible to diagnose at resolutions required to accurately model storms responsible for extreme precipitation. At exascale, climate simulations will produce results at fine enough resolution to investigate these relationships. However, the resulting data from these simulations will be far too large to save for post-simulation analysis. *We advocate for fitting statistical models inside the simulations as they run, a context known as in situ, which will facilitate scientific investigations using the full fine-scale data stream.* Figure 1 shows an example of the type of model we could consider, a Bayesian hierarchical spatial regression model. Precipitation extremes at each grid cell are modeled using extreme value distributions. Since extremes are rare, fitting models to individual grid cells can result in high variance and poor estimates. Instead, the model can be made more robust by smoothing the parameters of the extreme value model across space. Additionally, the parameters themselves can be functionally linked to other variables elsewhere in the simulation. Thus, we can use the fine-scale data to build more robust models for extremes that link extreme behavior to other climate patterns.

Rationale

The statistical and computational tools for fitting such models are currently missing. In situ analysis is a growing area of research, but the current focus is on simple statistical approaches mostly aimed at data reduction and often for visualization. Data reduction requires an inevitable coarsening that will make it difficult to investigate fine-scale relationships.

Models like the one in Figure 1 can be used to investigate these relationships, but fitting them in situ presents unaddressed challenges. First, the spatial component of these models typically requires estimating large correlation matrices that describe the correlation of every point with every point. This is computationally challenging. Second, Markov chain Monte Carlo is the gold standard for Bayesian approximation accuracy, but is an inherently serial algorithm that cannot scale to distributed computational resources. Third, the data in such simulations are also distributed in spatially contiguous manner (i.e. the data for a particular region is all contained on a single node), whereas many distributed estimation algorithms assume statistically independent batches of data or simply access to the entire data set. Finally, the simulation data must be considered in a streaming fashion: each temporal slice of data must be ingested as the simulation progresses because it is removed from memory as the simulation progresses.

Narrative

In order to do statistical analyses such as extreme value regression for precipitation, the above difficulties must be addressed. The time is ripe for such work as exascale computing will soon be available for DOE use and a number of raw ingredients are being developed in the statistics and computer science literature.

The core of a statistical approach like the one described above is a probabilistic model for spatial data. This can be used to model simulation data directly, e.g. a spatial model fit to monthly precipitation extremes. However, it is more helpful in the hierarchical structure described above where it can be used to do things like describe how parameters of a distribution can vary across space. This is especially powerful in the context of precipitation extremes which are rare in any given grid cell or small area. The spatial model, by building in the idea that neighboring regions should be similar, can use all of the precipitation data in a region to improve the estimation of the extreme value parameters in each grid cell.

A spatial model of this type will need to use sparsity where possible in order to be tractable in terms of computations and storage. It will also need to flexibly model correlation structure that changes across space and time. Deep, sparse Gaussian processes offer these advantages. Sparse Gaussian processes achieve tractability through the use of a parsimonious set of pseudo-data that to represent the large, complete set. In the case of precipitation extremes, this pseudo-data would consist of a small (relative to the size of the simulation data) set of locations and extreme value parameters for precipitation that can be used to interpolate and predict the extreme value parameters at any desired location. Deep Gaussian

processes can flexibly model different spatial correlation structures in different spatial regions. For example, this would allow the spatial correlation of precipitation extremes to differ between coastal and inland regions, but in smoothly varying manner. These models will need to be extended to handle the distributed and streaming nature of in situ processing. Scalability can be improved by using Krylov and randomized linear algebra methods.

Estimation for Bayesian models will also need improvements. Variational inference is well suited to the in situ setting. In variational inference, complicated Bayesian posterior distributions are approximated using a simpler, but still descriptive class of functions (e.g. multivariate Gaussians). The optimal distribution within the class is estimated by using an optimization approach instead of the sampling approach of Markov chain Monte Carlo. Often, this optimization uses stochastic approaches to gradient descent. Because this approach breaks free from the serial structure of most sampling algorithms, it should be extendable to parallel computing resources. There is significant work on applying variational inference to settings in which the data arrives sequentially in batches, but further work is needed to make these algorithms perform well in the in situ setting in which temporal slices of data arrive in batches of size one. Machine learning approaches can aid in this task in a number of ways such as learning summaries of the data (e.g. spatial correlation lengths for storm events or simple functional descriptions of storm probabilities) that can be used to fit statistical models and by predicting the results of simple parts of the statistical estimation.

Finally, implementation of the modeling and estimation is important. There are obviously many questions about the water cycle beyond precipitation extremes and many questions in climate beyond those about the water cycle. The modeling and estimation schemes should be released in a toolbox of reusable components that can be used to construct new models for new problems. This is a challenge in the in situ environment. The toolbox should be easy to use for data and climate scientists without needing to program in C or Fortran which are not conducive to data science. Nevertheless, the implementation of the statistical modeling should not be a bottleneck so it should match the speed of the Fortran or C that underlies the simulation itself. Finally, such a toolbox will need to be easily connected to simulations in these languages, including the accessing of data in memory without the need for copying. Implementation in Python is an obvious choice due to its ubiquity. However, we advocate for the use of the Julia language, which is growing in popularity for precisely the reasons we describe: it is both easy for data science programming and can be as fast as the compiled languages it will be embedded within.

Many of these issues are under study as part of ongoing Laboratory Directed Research & Development project at Los Alamos National Laboratory. The project is using the DOE's Energy Exascale Earth System Model as a development testbed. Although still in the early stages, this project is demonstrating the feasibility of the in situ approach to analyzing climate simulations.